XP 000198030

P.D. 02-10-'94
p. 356-364 = 9

H04N7/15M

On Multipoint Control Units for Videoconferencing

M. H. Willebeek-LeMair, D. D. Kandlur, and Z.-Y. Shae

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA.
mwlm, kandlur, zshae@watson.ibm.com

Abstract

This paper examines the issues involved in the design of conference servers that support multiparty, multimedia conferences. These servers, called Multipoint Control Units (MCUs) in the telephony world, coordinate the distribution of audio, video, and data streams amongst the multiple participants in a videoconference. The MCU is responsible for the processing of video and audio so that a conference participant can hear and see one or more of the other participants in the conference. It is also responsible for handling and forwarding the data streams from the participants. This paper presents different approaches to the design of an MCU to implement these functions. It also describes the design of a related device – a transcoding gateway that enables conferencing between participants using different video/audio equipment.

1 Introduction

In recent years, with the emergence of improved communication technologies with wider coverage and accessibility, videoconferencing has become one of the major new growth applications. Videoconferencing standards are being developed and more and more videoconferencing products are appearing in the market.

Videoconferencing solutions are currently evolving from several directions. On the one side, there are the circuit-switched (e.g., Narrowband ISDN or the Switched-56Kbps phone lines) types of solutions, which being motivated by the telephony industry, can be likened to it. On the other side are the packet-based network (e.g., Ethernet and Token Ring legacy LANs) solutions, which are designed to carry real-time traffic over existing computer communications networks. The advent of ATM [1] (Asynchronous Transfer Mode) might eventually allow these two approaches to converge.

Due to the stringent bandwidth, delay, and jitter requirements of real-time audio and video data, the solutions for the circuit switched and packet-based networks differ considerably. These differences include the encoder/decoder (CODEC) technology for video compression and decompression, the methods used to guarantee network performance, and in the provisions within the end-stations to handle the real-time traffic.

Videoconferences may be point-to-point or multipoint.

Point-to-Point. In a point-to-point videoconferencing a user is able to connect to only one other participant and communicate via video, audio, and shared data applications.

Multi-Point. A multi-point conference involves more than two participants and multimedia data is multicast from each participant to all others.

Each of the above scenarios involves the integrated communication of video, audio, graphics, and text data.

Approaches used to support multipoint conferences may be categorised as either distributed or centralised. In a distributed approach each end-station receives the video and audio streams from all, or some, of the participating end-station sources in the conference. Each end-station then composes these multiple incoming streams as desired. This approach is advantageous since it allows more flexibility and control at each end-station and minimises the distance that streams need to travel between source and destination. It requires additional processing capability

0-8186-6680-3/94 $04.00 © 1994 IEEE

BEST AVAILABLE COPY

at the end-station, and potentially greater bandwidth requirements from the network.

In a centralised approach, all conference streams are transmitted to a central server which then distributes composed and/or selected streams to the various participants. This central server is known as a Multipoint Control Unit (MCU). The centralised approach is advantageous since the complexity of handling multiple streams can be confined within this single shared device and the network bandwidth requirements can potentially be reduced. However, it reduces the flexibility of the end-station control, and may increase the latency of the real-time traffic since it has to go through an intermediate hop between source and destination.

This paper focuses on the centralized MCU approach, which is applicable to both computer network environments as well as circuit switched environments like telephony and ISDN. In addition to the multipoint conference control the discussion is extended to address the interoperability issue of different types of videoconferencing environments. We discuss some of the technical issues, standards activities, and possible solutions.

The rest of the paper is organized as follows. In Section 2, we describe what a Multipoint Control Unit (MCU) is and describe several different configurations with different functionalities. In Section 3, we describe the MCU functions as defined by the CCITT Standards for ISDN. The interoperability of heterogeneous conferencing environments is addressed in Section 4 with a discussion on Transcoding Gateways. A summary is provided in Section 5.

2 Multipoint Control Units

A Multipoint Control Unit (MCU) is a central server used to coordinate and distribute video and audio streams amongst multiple participants in a videoconference. The functions performed by an MCU are many and range in complexity. Basic MCU designs do little more than select a video stream from multiple incoming streams to be distributed to all conference participants. More complex designs perform a composition of multiple video streams before distributing the video stream to various participants.

To illustrate the differences between various MCU capabilities along with the tradeoffs between the complexity of these functions, the four MCU configurations shown in Fig. 1 are discussed.

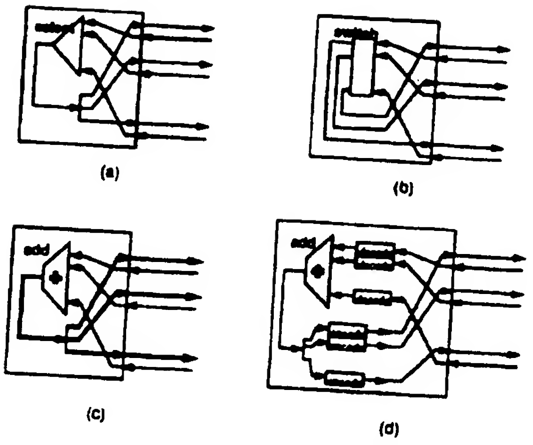

Figure 1: Multipoint Control Unit Configurations. (a) Single video select and broadcast. (b) Destination video select. (c) Video composition in the compressed domain. (d) Video composition in the pixel domain.

Select and Broadcast. The configuration in Fig. 1(a) is a simple design which selects one of multiple incoming streams to be broadcast to all participants. The criteria used to select the video stream to be distributed may vary. In a user controlled system, a designated user (conference manager) may send control information to the MCU to select the video source. Automatic systems may select the video source which corresponds to the active voice signal. The sensitivity of the selector must be adjusted to prevent an unintelligible thrashing between different sources. Synchronisation is required when switching between sources to insure that the switching occurs at frame boundaries and, in some cases, that a constant bit rate is maintained.

Destination Select. The configuration in Fig. 1(b) is slightly more sophisticated in that each recipient may independently select which video stream to receive. This requires a switch within the MCU with multicast capability. Signalling between each destination and the MCU is required to implement the selection. As with the *Select and Broadcast* configuration, switching between video streams must be synchronised to frame boundaries.

Compressed Composition. The configuration in Fig. 1(c) performs a composition of the incoming video streams in the compressed domain. Techniques for performing this composition of compressed video are presented in [2]. The resulting video stream bandwidth may be a multiple of the incoming bandwidth. For example, if three incoming streams of bandwidth β are composed into a single stream, where there is no overlap of the various sources in the composite, the resulting stream will have an average bandwidth of 3β. Hence, each station's incoming and outgoing link bandwidths are asymmetric.

Some video encoding algorithms consist of different types of frames, where some are *reference* frames which are encoded independent of any other frames, and *predicted* frames which are encoded based on reference frames. The decoding engine treats these frames differently, and cannot decode a predicted frame without the reference frames to which it applies. For these video streams, in order for the receiving decoder to be able to handle the composite stream as a single stream, the reference frames of each source need be synchronized to be mixed in the same composite frame, and switching should occur at reference frame boundaries.

Uncompressed Composition. The final configuration shown in Fig 1(d) also performs a composition of the incoming video streams, but does so in the pixel domain. This requires that each incoming stream be decoded and that each outgoing stream be encoded. Separate encoders for each outgoing stream are depicted in the figure due to clock synchronization requirements between the sender and receiver. If the synchronization requirement can be solved otherwise, a single encoder might suffice. By composing the incoming streams in the pixel domain, the resulting stream bandwidth can be adjusted to match different link speeds.

Another level of complexity may be added to the composition configurations by allowing each destination to select their own composition.

3 MCUs and ISDN

The International Telecommunication Union (ITU, formerly CCITT has defined several standards related to audiovisual services over circuit switched public networks (the Narrowband ISDN). These recommendations relate to the compression of audio and video, multiplexing (framing) of the multiple data types onto a single channel, and higher level signalling and control services. The H.320 document [3] gives an overview of services and the various associated standards. Amongst these documents, the H.231 standard [4] defines the functions of an MCU in the Narrowband ISDN environment. It classifies the functions as mandatory, to be supported by all compliant MCUs, and optional. The functions defined address

- Framing. On the input side, the MCU must demultiplex the incoming H.221 data [5] into audio, video, and application data. On the output side, after appropriate processing of the individual elements, it must regenerate an H.221 multiplexed stream for each connection.
- Audio processing. Audio processing is mandatory. However, the selection of functions provided can vary in range from simple format conversion and simple mixing, to selective mixing and private chats.
- Video processing. Video processing is considered optional, but most MCUs would be expected to provide some video processing functions.
- Data processing. Data processing is optional. Moreover, a meaningful exchange of data between disparate applications requires a standardization of higher level protocols for collaborative applications.

3.1 Framing

The MCU is required to support the multiplexing and demultiplexing of different data types using the H.221 standard. In addition to data, there are several control signals defined in the H.221 frame. Moreover, the values of some control signals are data as well as terminal dependent. Since the incoming data and terminal capability may be different from the outgoing one, the MCU must decode and stripe the incoming control signals and generate the appropriate outgoing control signals. We discuss the significance of some of the control signals below.

H.221 defines the frame structure for audiovisual teleservices using one or more B (64 Kbps) or H0 (384 Kbps) channels or a single H11 or H12 channel. This frame structure is used for several purposes such as: synchronization of changes in configuration, error recovery, and synchronization of the multiple B or H0 connections between the terminals. If a communication channel uses more than one B connections (the required data rate is more than 64Kbits/sec), it is possible that these multiple B connections are not aligned with each other. As such, the MCU has to provide the memory space to buffer the individual B connections and make use of the frame alignment signal (FAS) defined in the H.221 frame structure to synchronise and align them.

In order to provide an end-to-end monitoring of quality for the connection, H.221 requires the source to generate and insert 4 bits of CRC codes into every other H.221 frame. An MCU, therefore has to examine and stripe the CRC bits of each individual incoming data stream. After the processing, it is also required to generate the CRC code and to report a CRC error (if any of the incoming streams has failed the CRC) in the outgoing frame.

The Bit-rate allocation signal (BAS) is used to transfer commands that describe the capability of a terminal. Some possible commands include audio coding formats, video coding formats and their data rate. This command is protected with a (16,8) double error correcting code. Since the BAS command may be different between the incoming and outgoing frame, an MCU should be able to interpret, execute, and stripe the incoming commands. It must also have the ability to convert the incoming commands to appropriate outgoing commands.

3.2 Audio processing

The MCU is required to accept audio data in a variety of formats, with different data rates, such as: G.711 - 64 Kbps (or 56 Kbps) PCM with A-law or μ-Law companding, G.721 - 32 Kbps ADPCM encoding, and G.722 - 7Khz high quality audio at 48, 56, or 64 Kbps. The MCU must be prepared to decode and mix these different streams, but it may also provide additional functions.

- Simple mixing. This mandatory function combines the audio signals from all the participants and produces an output audio signal that is distributed to the participants.

- Selective mixing. The simple mixing function does not scale well as the number of participants in the conference increases. The inaudible signals from a large set of participants may add up to create a significant disturbance in the output signal. To prevent this, the MCU can implement a selective mixing function and limit the number of input streams that are mixed. The input streams are monitored and selected for the mixing based on the signal level (a form of silence detection). This enhanced mixing function is, for the most part, transparent to the end-stations. It may be enhanced by user inputs to permanently select some input streams - the conference chair being a possible candidate.

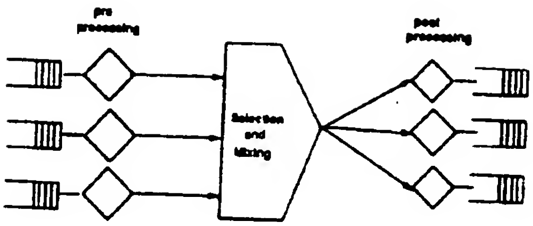

Figure 2: Audio processing functions.

- Private connections. Upon request, the MCU may provide a direct connection between two parties in the conference to facilitate a private conversation that is separate from the main discussion. Although this function is easily supported, it requires additional signalling/cueing between the end stations and the MCU.

It is possible to implement most of these functions directly on the digital audio streams. Figure 2 shows the different stages of audio processing: input pre-processing, selection, mixing, and output post-processing. The input pre-processing consists of expanding the coded audio samples into linear samples. It is noted that the different input streams may be coded using different standards and the processing is input dependent. This computation may be a simple table lookup, as in the case of expansion of μ-law or A-law samples, or a more complex decompression of ADPCM samples. The selection process consists of threshold testing the linear samples to detect and discard silent channels. Given this pre-processing, the audio mixing process is then a simple addition of the linear samples. The final post-processing phase involves re-compression of the mixed output to match the output required for each connection.

This process may be optimized in certain special cases, as in the case when all inputs follow the same standard, to eliminate the input and output processing and perform mixing directly on the input streams. This technique has the potential of reducing the processing power required for the audio processing, but may result in some degradation in audio quality.

3.3 Video processing

The H.261 standard [6] specifies the structure of the video stream and it ensures interoperability between video codecs from different vendors. It defines a constant bit-rate video stream whose bandwidth requirement is of the form $p \times 64$ Kbps, where p is an integer in the range 1..31. Two frame sizes are specified: the Common Image Format (CIF) which contains 352 × 288 pixels, and QCIF (quarter CIF) which contains 176 × 144 pixels. The video frame rate can vary between 7 and 30 frames per second. The standard specifies a canonical video stream as one which can be decoded by a *reference* decoder without causing any buffer overruns in the decoder.

In order to preserve the quality of the videoconference, it is necessary to ensure that the video processing does not increase the end–to–end latency. Moreover, the video processing and audio processing must be tuned such that the audio–video synchronisation in the original signal is maintained.

Selection. The MCU selects one of the incoming video streams and transmits it to all stations in the conference. The selection may be automatic (voice activated) or under the control of the conference director.

The selection function can be achieved without the need for video decompression and re-compression, within certain limits. In the steady state, the video processor copies video data from the selected input to all outputs. Since the input video stream is a compliant stream, the output seen by the receiving stations is also a compliant stream. However, the process of switching from one video stream to another poses some problems since the motion vectors in the new output stream could be inadvertantly applied to the pixels of the old output stream. A possible technique that can be employed to accomplish the switching is described below.

- The MCU requests the selected source to transmit a *refresh* frame.
- It monitors the old source to detect the end of the current frame. After the end of the current frame, it starts inserting dummy frames. (Alternatively, it can use the "freeze frame" signal to stall the output side.)
- When the requested refresh frame arrives from the new source the MCU starts sending the new frame.

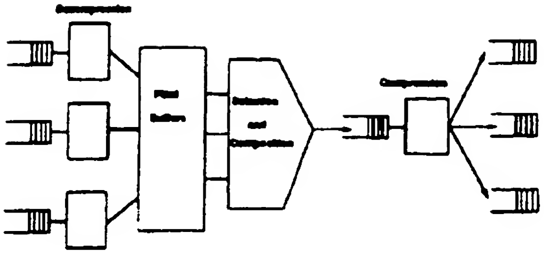

Figure 3: Video processing functions.

The video input data must be buffered in the MCU so that the delay in the video path can be matched to the delay in the audio path.

Mixing. Video selection has the limitation that the user is able to see only one person in the multi-party conference. Video mixing (or composition) may be used to provide the user with video images of more than one participant. For example, a 4-way video composition can be used to display four participants in a single video stream. Moreover, it is not necessary to modify the end station in any form since the MCU produces a single composite stream.

In the general form, the mixing process consists of the following phases: decompression, composition, and compression (see Figure 3). In order to implement an n-way composition, it requires n decompression units, a pixel composition unit, and one compression unit.

In some special cases, it is possible to construct a video processor which requires only one coder/decoder. Consider the situation in which the conference participants use QCIF images for the video. The MCU can implement the video composition in the following manner (see Figure 4).

1. Select a subset of the input sources (at most four)
2. Tile the QCIF inputs to form a CIF image
3. Decompress the compressed CIF image into pixel data

4. Subsample the CIF pixel data to produce a QCIF input

5. Compress the QCIF input and distribute to the outputs

This technique substantially reduces the complexity of the video composition process. It composes the input QCIF images in compressed form, so the amount of data that it has to process is substantially smaller. Moreover, the decompressed pixel data forms a contiguous region in the pixel buffer, making it easy to apply the subsampling to generate the output QCIF image. An additional benefit of this approach is that there is now only one input device to the pixel buffer which greatly simplifies the design of the buffer arbitration mechanism.

Using this technique it is possible to create a low-cost implementation of the video mixing function for a very important special case. It is possible to provide 4-way video composition for video conferencing with basic rate ISDN (128 Kbps)[1].

It is noted that although the video mixing function provides users with multiple video images, it has some significant drawbacks. The decompression and compression stages add a significant delay in the path, so the end-to-end delay is increased. Also, in order to maintain audio-video synchronisation it is necessary to delay the audio stream in the MCU.

3.4 Data processing

The data stream carried in the ISDN channel is dependent upon the higher-level protocols that control the videoconference. These protocols are currently being defined by the ITU (as the T-series standards).

4 Advanced Functions

The videoconferencing solutions for circuit switched and packet-based networks may differ considerably [7]. Circuit switched solutions have the advantage that the end-to-end network delay can be minimised and bandwidth can be guaranteed. However, compressed video is inherently variable bit rate and special provisions must be made to force the compressed video stream to conform to a constant bandwidth. This results in larger delays in the video coder/decoder (CODEC) element. A rate absorption buffer located at the output of the compression engine can be used to provide feedback to the compression engine to vary the compression algorithm to generate a constant bit rate. This rate absorption buffer must be large enough to provide adequate feedback, and as a consequence adds latency to the compression entity.

Packet-based solutions are advantageous since they can more naturally support the variable bit rate and reduce the CODEC delay. On the other hand, these solutions must be designed to accommodate larger network delays and jitter. Furthermore, packet-based LAN solutions which operate over high speed networks can cost effectively be designed for higher quality, higher bandwidth, video [8]. For the wide area, however, it may be more cost effective to use lower bandwidth solutions designed for circuit switched or phone line connections.

It can be seen that the encoding algorithms for the video compression tend to strongly depend upon the network environment. Numerous alternative standard compression algorithms (e.g., Motion-JPEG, MPEG, Indeo, etc.) as well as proprietary algorithms are being used for videoconferencing. Furthermore, new compression algorithms are being developed for LAN-based PC videoconferencing [9] and ATM (Asynchronous Transfer Mode)[1]. The result is a heterogeneous environment of compression algorithms which can be likened to the network transport protocol environment. It would be desirable to enhance the functionality of the MCU to be able to translate between different encoding formats to enable interoperability between different videoconferencing environments. We term an MCU with this functionality a *Transcoding Gateway*.

A block diagram of a Transcoding Gateway (TCG) is illustrated in Fig. 5. In this example, videoconference stream A is connected to videoconference stream B. The conference streams from A and B enter the TCG via ports A_{in} and B_{in}, respectively. The conference streams to A and B exit the TCG via ports A_{out} and B_{out}, respectively. Hence, conference streams through A_{in} and A_{out} conform to the same protocol (Protocol-A), as do the streams through B_{in} and B_{out} (Protocol-B). However, Protocol-A and Protocol-B may differ dramatically. The node processor (NP) translates the control information from Protocol-A to Protocol-B and vice-versa. Furthermore, the data[2]

[1] QCIF images are used extensively for video conferencing with basic rate ISDN (128 Kbps).

[2] Here data refers to the encoded video, audio, or data that is being used as part of the videoconference application.

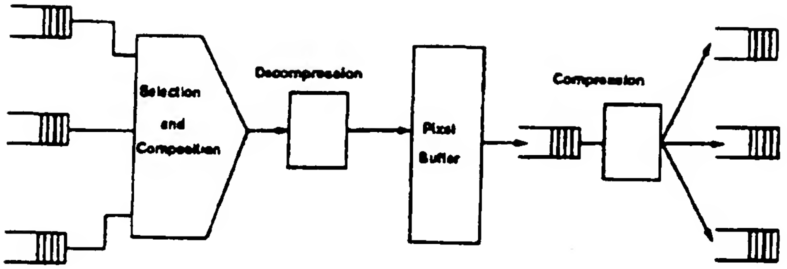

Figure 4: Compressed video mixing.

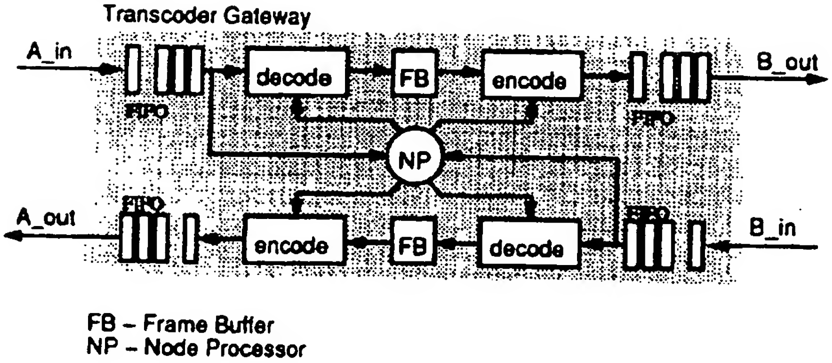

Figure 5: Gateway Transcoder block diagram.

stream from each input port must be decoded to a common (raw) format before it can be re-encoded according to the format required at the destination output port. A frame buffer is located between the decode and encode engines where data is stored in its raw form.

In a more general model, each incoming stream could be sent out as multiple outgoing streams using different encoding formats. It would also be possible to receive multiple incoming streams conforming to different encoding formats and compose them into an integrated outgoing steam.

Besides the encoding translation, there are other changes between an input and output stream that can be made. The video and audio data is, typically, in compressed form and unless a direct translation between two different compression algorithms is possible, the incoming stream must first be decompressed to a raw format and re-compressed by the new algorithm before being retransmitted. Other translations such as frame sise. frame rate, or compression ratio may also be performed. These translations may be required due to end-station or network capabilities.

4.1 Sample Transcoding Scenario

To achieve the benefits of high quality across the LAN, and low-cost across the WAN, the result is a heterogeneous environment as shown in Fig. 6. As an example, the LAN solution might use Motion-JPEG video compression over UDP/IP and the WAN connection could be H.320 with H.261 video compression over ISDN. The connectivity between the LAN and the WAN is through a TCG. The TCG would have to translate the connection setup signalling between both environments as well as any other signalling that occurs during the conference. Furthermore, the TCG would have to translate the packetized M-JPEG stream to a framed H.261 video stream, and vice-versa. The ISDN H.320 environment could view the TCG as an H.320 end-station or, if more sophisticated, the TCG could behave as an MCU

The concept of multiple MCUs is quite interesting, and can be extended to consider hierarchical MCU configurations. As food for thought, Fig. 6 illustrates a scenario with three MCUs: one in each of the LAN environments, and one in the ISDN cloud. In the simplest scheme each MCU gathers streams from all stations within its environment and treats the other MCUs as another end-station. In the configuration shown in Fig. 6, MCU_1 may not be aware of MCU_3, only of MCU_2. In fact only MCU_2 is aware of both other MCUs, but treats them as end stations. Clearly, more complex and interesting scenarios are possible.

5 Summary

In this paper we have described the functions required to support multimedia, multiparty videoconferences. These functions may be implemented in a decentralized manner in each participating end-station, or in a centralised manner as a multipoint control unit. We have described various alternative approaches to providing the video and audio processing functions in an MCU, with particular emphasis on the narrowband ISDN environment. We have also presented the case for transcoding gateways between the public (ISDN) networks and private (packet switched) data networks.

References

[1] J. Y. L. Boudec, "The asynchronous transfer mode: a tutorial," *Computer Networks and ISDN Systems*, vol. 24, pp. 279–309, May 1992.

[2] Z.-Y. Shae and M.-S. Chen, "Mixing and playback of JPEG compressed packet videos," Research Report RC 16068, IBM Research Division, T. J. Watson Research Center, Yorktown Heights, NY 10598, Aug. 1990.

[3] "Narrow-band visual telephone systems and terminal equipment." CCITT Recommendation H.320, July 1990.

[4] "Multipoint control units for audiovisual systems using digital channels up to 2 mbit/s." CCITT Recommendation H.231 (draft), May 1992.

[5] "Frame structure for a 64 to 1920 kbit/s channel in audiovisual teleservices." CCITT Recommendation H.221, July 1990.

[6] "Video codec for audiovisual services at p x 64 Kbit/s." CCITT Recommendation H.261, July 1990.

[7] M. Willebeek-Lemair, F. Schaffa, and B. Patel, "End-to-end multimedia communication," Research Report RC 18454, IBM T. J. Watson Research Center, Oct. 1992.

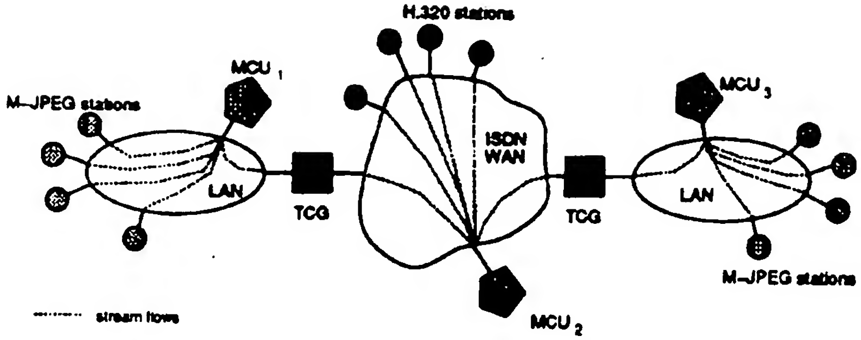

Figure 6: Heterogeneous videoconferencing environment.

[8] M.-S. Chen, Z.-Y. Shae, D. D. Kandlur, T. P. Barsilai, and H. M. Vin, "A multimedia desktop collaboration system," in *Proceedings GLOBECOM 92*, IEEE, Dec. 1992.

[9] "Generic coding of moving pictures and associated audio." ISO/IEC Recommendation H.26x, working draft, Sept. 1993.

THIS PAGE BLANK (USPTO)

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

□ BLACK BORDERS

□ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES

□ FADED TEXT OR DRAWING

□ BLURRED OR ILLEGIBLE TEXT OR DRAWING

□ SKEWED/SLANTED IMAGES

□ COLOR OR BLACK AND WHITE PHOTOGRAPHS

□ GRAY SCALE DOCUMENTS

□ LINES OR MARKS ON ORIGINAL DOCUMENT

□ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY

□ OTHER: ______________________________

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

HIS PAGE BLANK (USPTO)